Architecture and Evolution of Blade Assembly in -propeller Lectins

Bonnardel, Francois; Kumar, Atal ; Wimmerova, Michaela ; Lahmann, Martina; Perez, Sergei; Varrot, Annabelle; Lisacek, Frédérique ; Imberty, Anne

Structure

DOI:
10.1016/j.str.2019.02.002

Published: 07/05/2019

Peer reviewed version

Architecture and evolution of blade assembly in β-propeller lectins

François Bonnardel [1,2,3], Atul Kumar [1,5], Michaela Wimmerova [5,6], Martina Lahmann [7], Serge Perez [8], Annabelle Varrot [1], Frédérique Lisacek [2,3,4]*, and Anne Imberty [1]*

1. Univ. Grenoble Alpes, CNRS, CERMAV, Grenoble, France.
2. Swiss Institute of Bioinformatics, Geneva, Switzerland.
3. Computer Science Department, UniGe, Switzerland.
4. Section of Biology, UniGe, Switzerland.
5. CEITEC, Masaryk University, Brno, Czech Republic
6. NCBR, Fac.Sci, Masaryk University, Brno, Czech Republic
7. School of Chemistry, University of Bangor, Bangor, United Kingdom,
8. Univ. Grenoble Alpes, CNRS, DPM, Grenoble, France.

* To whom correspondence should be addressed. Anne Imberty (anne.imberty@cermav.cnrs.fr, Tel: +33 476 03 76 40) Frédérique Lisacek (: frederique.lisacek@isb-sib.ch, Tel: +4122 379 58 98)

Abstract

Lectins with a β-propeller fold are proteins remarkable by their ability to bind glycans on the cell surface through multivalent binding sites and appropriate directionality. Since these propellers are formed by a repeat of short domains, they are of high interest as a result of evolutionary duplication. Such repeats are difficult to identify in translated genomes and usually not correctly annotated in sequence databases. To address these issues, we defined the blade signature of the five types of β-propeller lectins using 3D-structural data. With these templates, we predicted 3887 β -propeller lectins in 1889 different species and organised this new information in a database that can be searched through a web interface. The data reveals a widespread distribution of different β-propeller lectins in the living kingdom. For example, some are present in pathogenic bacteria and represent interesting targets for anti-infectious therapeutic strategies. Prediction also emphasised different architectures, and association with other proteins. Interestingly, we uncovered a novel scenario to create a β-propeller, by assembling two short proteins with 3 blade repeats. To confirm the hypotheses, a predicted protein coded in the genome of a fresh water bacterium, *Kordia zhangzhouensis*, was produced and characterized. The crystal structure confirms a new intermediate in the evolution of β-propeller assembly and demonstrates that our software and database are excellent tools for the identification of novel β -propellers.

Introduction

Among the players in glycobiology, lectins are protein receptors that can bind at least one carbohydrate, and with no enzymatic function[1]. Lectins are generally multivalent and such multiplicity of carbohydrate binding sites favours the strong avidity to glycoconjugates available in multiple copies on all cell surfaces. Lectins are involved in a range of biological processes taking place between cells. For example, they participate in the interaction between microorganisms and hosts cells (pathogenicity, symbiosis…). Despite such a prevalent role, lectins are rather poorly characterised in protein databases. To overcome this shortcoming, we launched the Unilectin3D database[2] that includes a large number of classified and manually curated lectin 3D-structures, with information on their fold, oligomeric structure and carbohydrate binding site(s). The Unilectin3D collection highlights the diversity of folds that lectins adopt, and the high frequency of the occurrence multimeric structures. However, for some lectins, multivalency is not created by oligomerization, but by tandem repeat of conserved carbohydrate binding domains. Such tandem repeats are observed in the so-called β-propeller lectins.

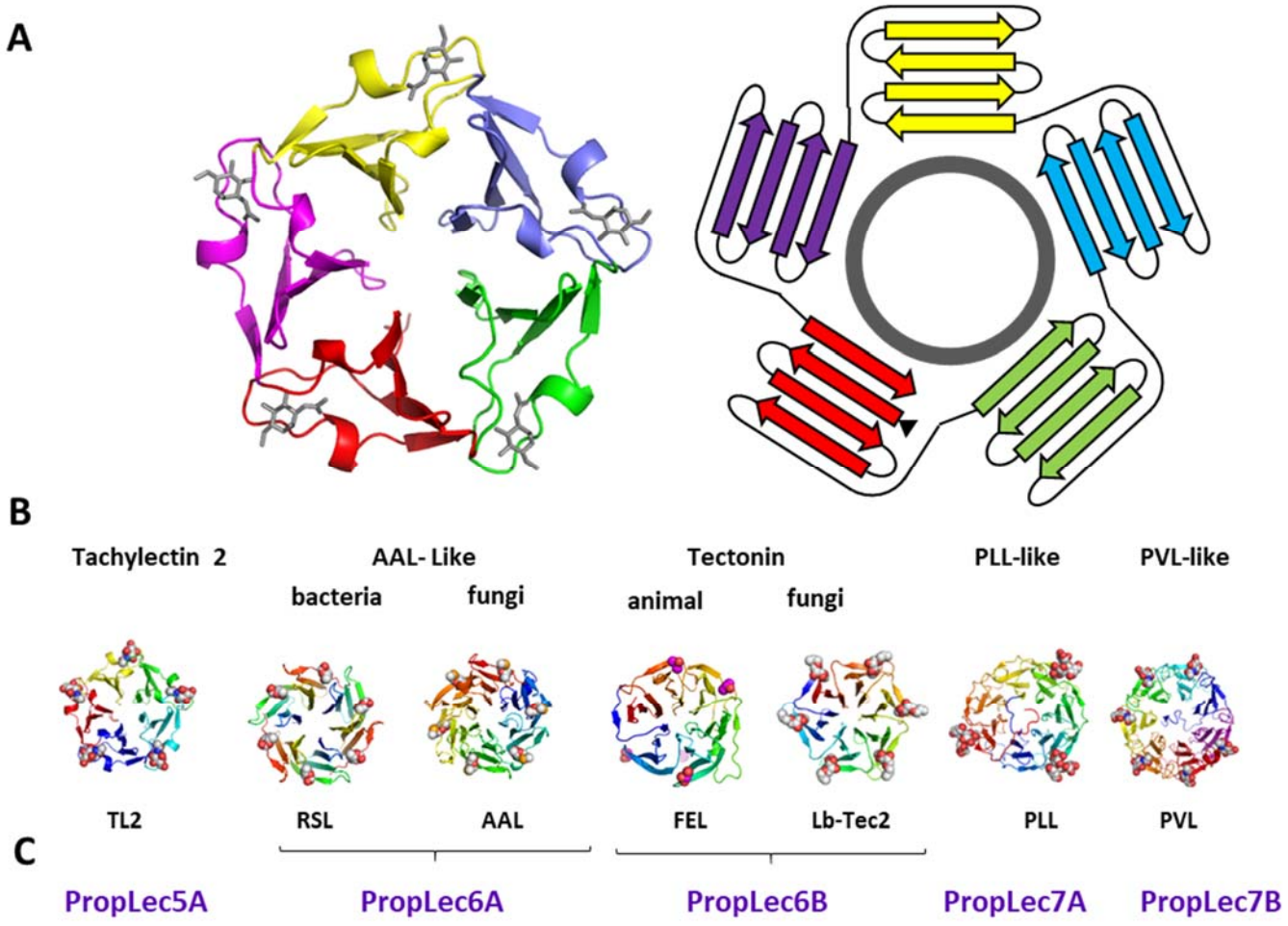

Figure 1. *A. Example of lectin β-propeller structure: the 5-bladed tachylectiy-2 (1TL2) complexed with 5 GlcNAc residues and its schematic representation. B. Structures of the seven classes of PropLecs in Unilectin3D (see Table S1 for details on each structure). C. Simplified nomenclature for the five families in the PropLec database.*

The β-propeller is a fold widely distributed in Nature[3, 4]. β-propeller proteins adopt a donut shape made of four to ten repeats (or blades) of four-stranded β-sheets[3-5] (Figure 1A). Their functions are broad, generally related to an enzymatic active site located in the centre of the structure. Although very variable in amino acid sequences, β-propellers have been proposed to derive from a single peptide through multiple episodes or duplication and diversification[6, 7]. The β-propeller fold is a very stable arrangement of repeats and allows for optimum presentation of multiple binding sites. Such topology is perfectly suited to bind carbohydrate epitopes on glycoconjugates presented on cell surfaces. It is therefore not surprising that this fold has successfully been adopted by nature for lectin functions. At the present time, Unilectin3D contains 52 X-ray structures from 13 different β-propellers proteins (PropLec) with five to seven blades that have been classified in seven different groups (Figure 1B).

Tachylectin-2, isolated from horseshoe crab, is the only 5-blade PropLec structurally characterized[8]. It binds to *N*-acetylglucosamine (GlcNAc), a glycan epitope present in the cell wall of pathogens, and is thought to be involved in the innate immunity of invertebrates[9]. *Aleuria aurantia* lectin (AAL) from orange peel mushroom is a 6-blade β-propeller[10] that binds to fucose (Fuc). AAL-like β-propellers have also been structurally crystallized from pathogenic fungi, such as *Aspergillus fumigatus*[11], where they play a role in eliciting host immune response[12, 13]. Bacteria such as the plant pathogen *Ralstonia solanacearum* and the human pathogen *Burkholderia ambifaria* produce lectins with high similarity to AAL but containing only 2-blades[14, 15]. These are the only known examples of natural β-propellers formed by oligomerisation, representing probably some ancestral form of the fold. Tectonin, a 6-blade β-propeller that binds to methylated monosaccharides generally associated with pathogens, has been structurally characterized[16]. It is present in fish (FEL), with a proposed role in the antibacterial protection of the eggs, as well as in the mushroom *Laccaria bicolor* (Lb-Tec2) [17]. In the latter case, four tectonins oligomerize in a virus-like shape that is involved in defence against worms feeding on mushrooms[17, 18] . The same anti-feeder role of the 7-blade PropLec in *Psathyrela velutina* (PVL) or *Agrocybe aegerita* (AAL2) mushrooms that bind GlcNAc[19, 20] is likely. A different 7-blade β-propeller has been characterized in two species of *Photorhabdu*s bacteria (PHL and PLL) with evidence for dual specificity for Fuc and galactose (Gal) in different binding sites[21, 22].

PropLecs are of high interest for their role in defence and self-immunity. Since some of them are involved in host-pathogen recognition, they are also promising targets for glycomimetic compounds that could present anti-infectious properties. Designing multivalent molecules that fit the specific binding sites arrangement of β-propellers in pathogenic micro-organisms has been key to obtain high-affinity inhibitors[23-25]. Because of their ability to bind strongly to glycoconjugates on cell surfaces, PropLecs are also useful biomarkers, for probing the glycosylation of proteins[26, 27], for labelling cancer cells[14], or as tools to study the dynamics of glycolipids in membranes[28]. Finally, PropLecs have been engineered, dissected in smaller pieces and reassembled to build artificial proteins for understanding stability and folding processes[29-31].

β-propeller structures are easily identified by their characteristic shape. As a result, β-propellers are in general well described in structure databases. For example, the CATH-GENE3D database[32] has categories for propellers from 3 to 8 blades, yet not all PropLecs are included. In fact, β-propeller lectins are difficult to identify based on their amino acid sequence. The presence of short repeated peptide motifs (30 to 50 amino acids) challenges classical search programs that are based on sequence alignment. This setback, in turn,

impacts the definition of family as well as the reliability of sequence-based genome mining tools. For example, the Pfam protein family database[33] defines family profiles based on domain similarity, but Pfam profiles matching PropLecs cover either part(s) of one blade (each blade is 46 to 58 amino acids long) or the whole propeller. As a result, no current tool can, as is, efficiently mine β-propellers, and they usually miss the conserved carbohydrate binding sites of PropLecs.

We developed here a precise method to detect automatically PropLecs in sequence databases. Robust peptide motifs corresponding to the repeating unit of each PropLec family were derived from the alignment of the blade sequences whose boundaries are delineated in the 3D structures. Conserved regions set the definition of family profiles and the HMMER profile search tool[34] was used to search for similar proteins in the non-redundant protein dataset from Uniprot /Uniref100[35]. The likelihood of predicted PropLecs is scored. This prediction tool can be used to identify new targets for antibacterial drugs, association between the carbohydrate-binding and enzymatic domains, and new protein oligomerization forms. The examination of predicted results led us to unveil an alternate scenario of blade assembly hitherto not observed. We validated this potential novel way to assemble β-propeller in a predicted PropLec of *Kordia zhangzhouensis* by solving the crystal structure of this new lectin.

Results

Definition of conserved motif in each β-propeller lectin family

The presence of repeated domains in PropLecs challenges their automatic detection in genomes. Our strategy was to turn this into an advantage by defining conserved motifs corresponding to the blade signature in each family, and then to search multiple and successive occurrences of these motifs in genomes. The seven sub-groups of PropLecs that are described in Unilectin3D were defined based on structural similarity and taxonomy. By focusing only on structural and sequence similarity, we reduced this number to five PropLec families (Figure 1C). To simplify the nomenclature, each family has been named according to the number of constituting blades, e.g. PropLec5A, PropLec6A, PropLec6B, PropLec7A and PropLec7B

The structural information in the 13 different PropLecs that have been crystallized so far was used to identify the blade signature of each PropLec family (Table S1 in supplemental information). The peptide sequences were first processed with the RADAR software [36] in order to align the repeated regions. This alignment was refined on the basis of 3D-structural information, which entailed the adjustment of repeat boundaries to the definition of blades. When necessary, alignments were shifted along the sequence so as to centre each blade on the 3D structure. The resulting blade sequence alignments are displayed in Supplementary information (Fig S1 to S5 in supplemental information). They served as the basis for determining conserved motifs and defining characteristic profiles in the form of Hidden Markov Models (HMM). These models were generated with the HMMbuild tool of the HMMER software suite [34]. HMM profiles identify similar domains depending on the amino acid frequencies at each position of the blade and on the amino acids in previous positions.

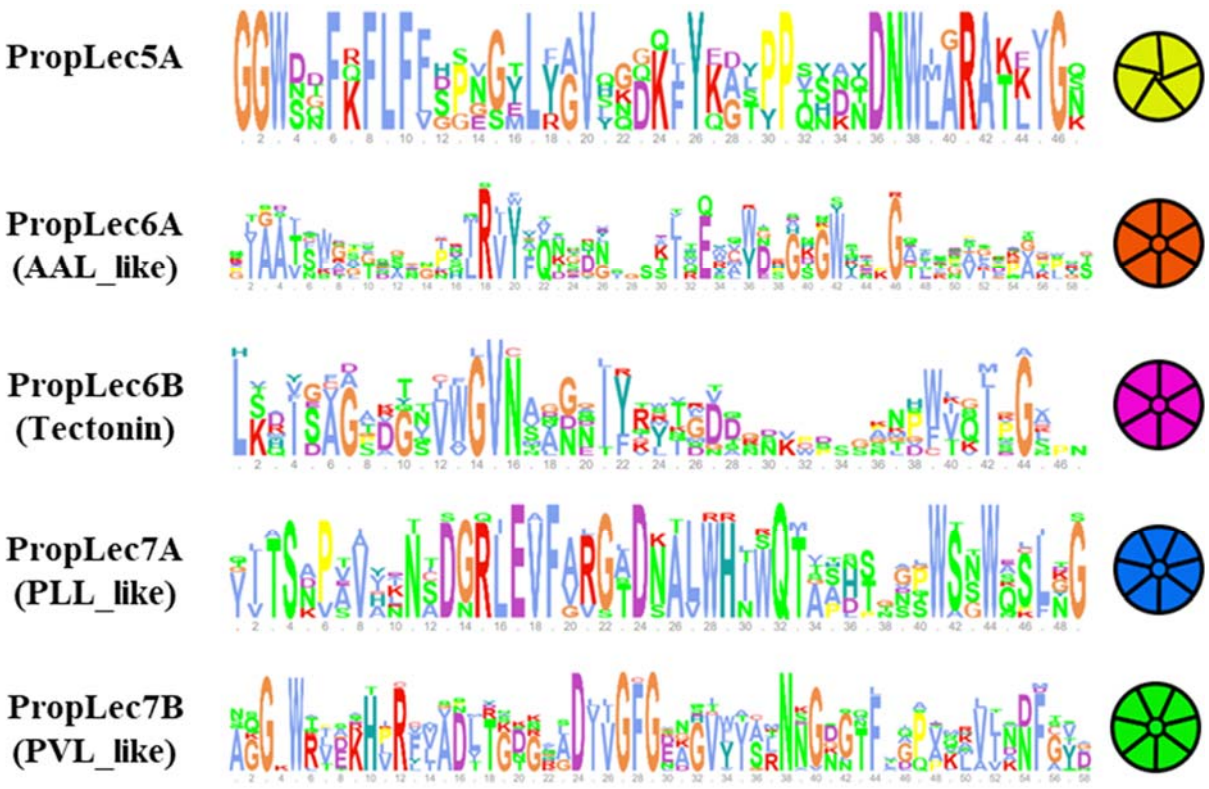

Figure 2. *Signature motif extracted from blade alignment for the 5-blade family of PropLec. Amino acids in one-letter code are coloured by class of properties, and the size of the letter corresponds to the frequency of the amino acid in the alignment. Complete sequence alignments are provided in Supp. Info (Figure S1 to Figure S5)*

As seen in Figure 2, each of the five PropLecs families have very different HMM motifs. Interestingly, the most conserved amino acids often correspond to the ones involved in the binding of the carbohydrate ligand, which indicates the conservation of function, in addition to structure.

The PropLec database

In order to identify PropLecs in other organisms, the designed motifs were fed into HMMSEARCH to process the UniRef100 non-redundant protein database (12/09/18 version containing 124 million distinct protein sequences). The predicted protein sequences were filtered with an e-value set to 0.01 while other parameters were left to default values. This search returned 3877 putative PropLec sequences containing a total of 20090 conserved blades domains (Figure S6).

A dedicated interface for mining the PropLec database is available at https://www.unilectin.eu/propeller/. For each predicted protein, information is displayed using an in-house sequence viewer and an amino acid conservation plot or sequence logo redeveloped with D3JS[37]. Both the reference and the predicted blades were aligned with the MUSCLE software[38]. The resulting multiple sequence alignment (MSA) is used to define

one score that evaluates the similarity of each blade with the defined reference motif (see Methods section and Figure S7). A key parameter for analysing results is the cut-off value for this score on the third quartile.

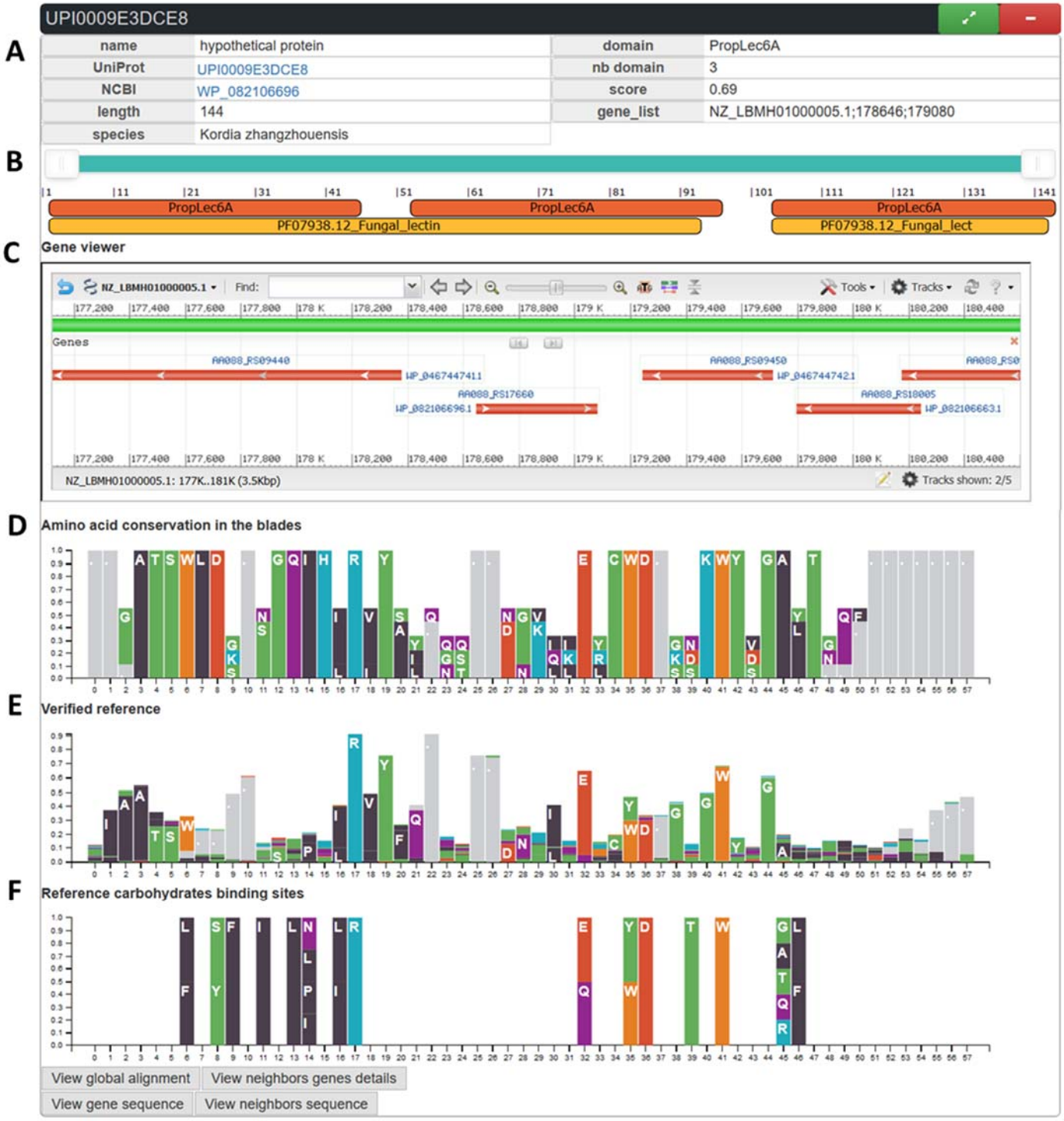

Figure 3. *Example of an entry for a predicted lectin sequence in the PropLec database. A. Information about the sequence and species. B. 2D sequence feature viewer with localisation of the predicted blades and potential Pfam domains, with a drag and drop button to zoom in on the sequence. C. NCBI viewer for the corresponding gene and chromosome D. Barchart of the amino acid conservation between blades of the predicted protein. E. Amino acid conservation of the reference blade. F Amino acids involved in carbohydrate recognition in the reference blade.*

The information stored in the database can then be searched using different filters, such as the similarity score mentioned above, but also information related to lectin families, sequence (number of blades..), biological origin (species..) and others. The search interface is displayed in Figure S6. By default, sequences are filtered using a minimum similarity score of 0.25 and synthetic genes or partial sequences are excluded. This filtering resulted in 3605 proteins of interest. Most of them are predicted to belong to families with 7-blades (54% for PropLec7A and 22% for PropLec7B). The two 6-blade families are evenly populated (13% for PropLec6B and 8% for PropLec6A) and less than 3% belongs to the 5-bladed tachylectin family (PropLec5A).

Searches in the database generate lists of sequences with information covering family, number of blades, score, sequence length and taxonomy. Each entry can then be expanded to a full page showing further information on gene and protein sequences with cross-links to external resources as well as details of the alignment and amino acid conservation in the form of histograms. The latter highlight the comparison of the family reference and predicted motifs and allow for a visual check equivalent to the similarity score. Furthermore, the amino acids involved in the carbohydrate-binding site of the reference lectin are singled out below the alignment, as an instant evaluation of the likelihood of a lectin function. Figure 3 exemplifies a sequence from the freshwater bacterium *Kordia zhangzhouensis*. Zooming in and out is made possible both for the in-house simplified protein viewer and the NCBI gene viewer[39]. Further information, such as the details of binding site contact with different carbohydrates (if available), the full alignments of all blades (predicted protein and reference) and details on neighboring genes are also visualized on the page.

Occurrence of PropLecs in the living kingdom

The distribution of PropLecs in the tree of life can be analysed through the interface, with both sunburst and tree representation available (Figure 4). No significant bias is observed for the two main branches with 75% PropLecs sequences of bacterial origin and 24% of eukaryotes (the non-redundant NCBI database reports 76% and 21% sequences from bacteria and eukaryotes, respectively[40]). Only 28 proteins have been predicted in Archae (0.7%) that appear to be under-represented. Interestingly, no PropLec sequence is identified in virus genomes with the exception of a synthetic one used in phage display[30] that has been therefore filtered out of the database. Bias occurs in eukaryote subgroups, with an over-representation of PropLecs in fungi genomes. As much as 11% of PropLecs are in fungi (404 proteins) while fungal sequences represent less than 3% of the RefSeq database. Plant genomes do not contain any PropLec sequence and they are rare in algae. Similarly, we could not identify any sequence in birds and mammals, which comforts the hypothesis that PropLecs play mainly a role in innate immunity that has been partially replaced by acquired immunity in more evolved organisms. It should be noted that putative PropLecs were proposed in human as members of the PropLec6B family and referred as leukolectin or hTectonins[41]. However, the sequence of human leukolectin (GenBank Accession: ACM77812) is 100% identical with the salmon tectonin. Searching the human genome (BLAT search on UCSC browser: https://genome.ucsc.edu/cgi-bin/hgBlat) with the leukolectin gene sequence did not return any hit. Altogether, these observations point to a probable contamination problem during RNA sequencing. The other putative human tectonins[41] have not been demonstrated to fold as β -propellers and they do not present any of the conserved motif that we identified.

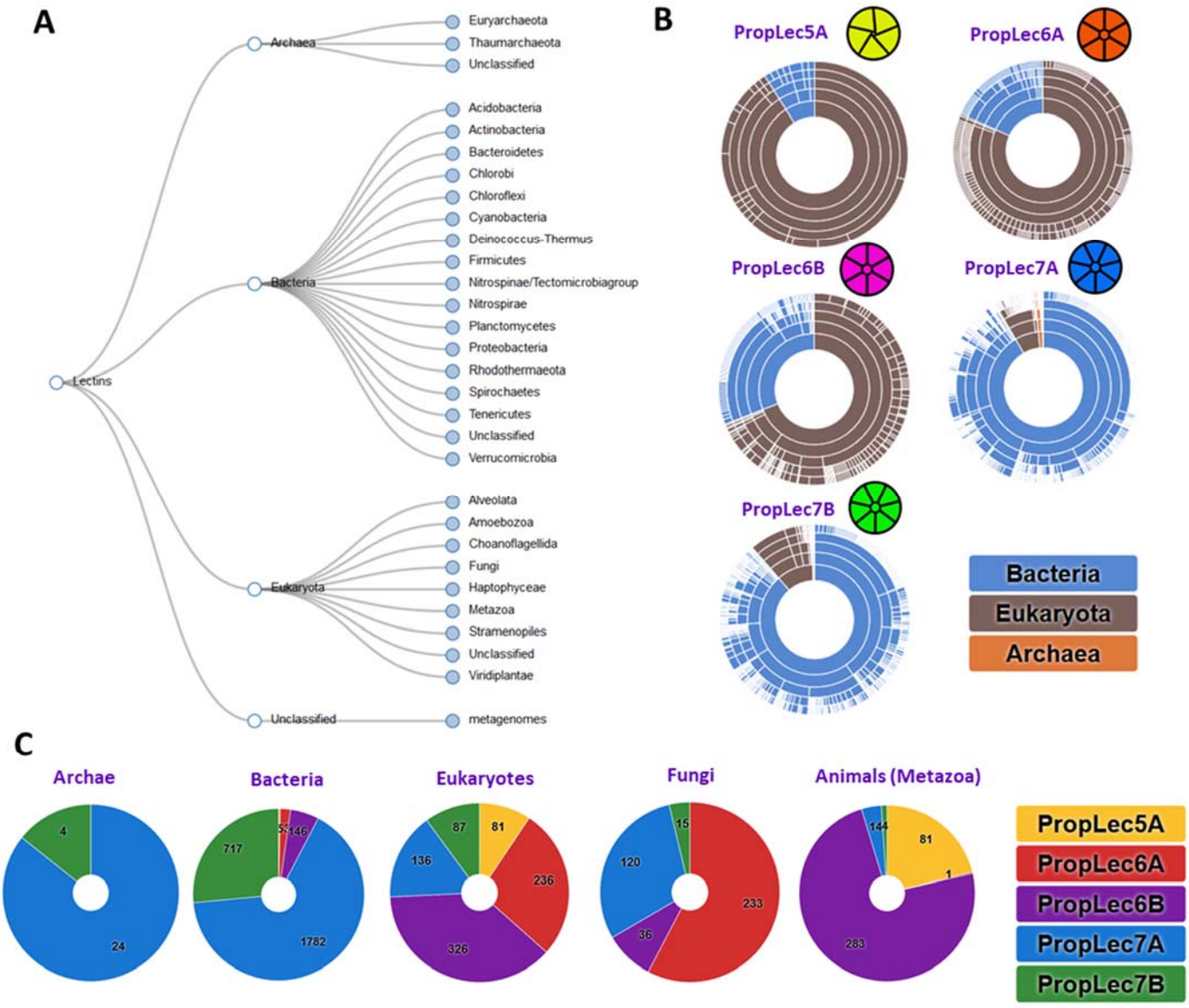

Figure 4. *Occurrence of PropLec sequence in genomes. A. Searchable tree in the PropLec database. B. Sunburst statistics for the origin in each PropLec family. C. Sunburst statistic for PropLec families in selected domains of life.*

The different families of PropLecs do not occur equally in Nature (Figure 4). All five families are present in bacteria and eukaryotes, albeit with very different populations. PropLec5A (tachylectin) is an exclusive animal lectin, identified in invertebrates (Cnidaria and crabs), xenops and fishes. Archaea and bacteria genomes contain mostly PropLec7A, the fucose/galactose lectin recently identified from several *Photorhabdus* species. Eukaryotes genomes contain all five families of PropLecs but the distribution is different in fungi, where a majority of PropLec6A (the AAL lectin) is observed, in contrast with animals, that contain mostly PropLec6B (tectonin).

Since many pathogens use lectins for recognition and adhesion to host tissues, such lectins are considered as targets for the development of anti-adhesive compounds[42, 43] and their identification may have a therapeutic relevance. A list of microorganisms that cause diseases in human is available from NIH NIAID Emerging Infectious Pathogens. Afilter in the main page of the database allows for selecting only PropLecs in such organisms. We identified PropLecs in more than 20 pathogenic microorganisms, and the ones that are more threatening for human health or characterized as emergent threats are listed in Table 1. Among them, only two lectins, AFL in *Aspergillus fumigatus* and BambL in *Burkholderia ambifaria,* have been fully characterized[11, 44]. AFL was demonstrated to be located on the fungal conidia and to play a role in host defence by interacting with the inflammation response[11, 12]. The lectins listed in Table 1 would therefore be of high interest for the understanding of pathogen-host interactions.

Table 1: Identification of PropLecs in the genomes of pathogenic micro-organisms.

	species	propfamily	disease	PMID
Gram+ bacteria	*Bacillus cereus*	PropLec7A	Food poisoning	23488744
	Clostridium botulinum	PropLec7A	Botulism, food poisoning	28800585
	C. tetani	PropLec7A	Tetanus	25638019
	Nocardia mikamii	PropLec7B	Opportunistic lung infection	19915112
Gram- bacteria	*Burkholderia ambifaria, B. cepacia*	PropLec6A	Opportunistic lung infection	22170069
	B.ubonensis	PropLec6A, PropLec6B	Opportunistic lung infection	27303639
	Coccidioides immitis	PropLec7A	“Valley fever”, meningitis	28597822
	Ralstonia pickettii	PropLec6A	Emerging nosocomial infection	16337309
Fungi	*Aspergillus fumigatus*	PropLec6A	Aspergillosis, lung infection	10194462
	Fonsecaea erecta	PropLec6A	Chromomycosis, skin infection	11204152
	Phialophora attae	PropLec6A, PropLec7A	Chromomycosis, skin infection	26586868
	Trichophyton tonsurans	PropLec6A, PropLec7A	Dermatophytosis, scalp infection	23053563
Oomycetes	*Pythium insidiosum*	PropLec6B	Pythiosis, multisystemic infection	20800978

Prediction of topology and modular associations

Since the family motifs have been defined to correspond to one blade length, the search procedure can predict the number of blades that are conserved in the sequences. In the database, the number of blades varies from 1

to 26, but most sequences are predicted to include 6 or 7 blades, in agreement with the known 3D-structures (Figure 5). A significant number of sequences show a lower number of blades than expected, such as 6 blades for PropLec7A and PropLec7A, which is explained by variation in amino acids in one blade of the protein (degeneration of sequences). Larger number of blades generally corresponds to the tandem repeat of several propellers in the sequence, explaining the highest occurrence for 12 and 18 blades, corresponding to 2 or 3 propellers in the same sequence.

Carbohydrate-binding domains or modules (CBM) are related to lectins, since they bind to carbohydrate, but they are usually monovalent. CBMs act as substrate binding and can be combined with carbohydrate-active enzymes [45]. It is therefore of interest to analyse the modular architecture of the predicted PropLecs to check if they could also associate with enzyme-active domains. The database interface has been designed to search for the occurrence of such modules. Twenty-nine distinct domains not overlapping with PropLecs domains were identified, some of them are listed in Table 2. Glycosyl hydrolases, or other enzymes acting on carbohydrates are often attached to PropLecs, which are then supposed to act as substrate recognition modules. Other enzymes are also identified such as peptidases or peroxidases. Interestingly, PropLec can also tandem with other carbohydrate-binding proteins, such as C-type lectins.

Table 2: Selection of functional domains identified with PropLecs with a modular design on the same peptide.

Family	**Architecture**	***Species***	**Associated protein**	**Pfam**
PropLec6A		*Aspergillus lentulus*	Aldo-keto reductase yakc	PF00248
PropLec6A		*Actinopolymorpha singaporensis*	Cysteine peptidase	PF00112
PropLec6B		*Branchiostoma belcheri* (lancelet)	C-type lectin	PF00059
PropLec6B		*Branchiostoma belcheri*	Animal haem peroxidase	PF03098
PropLec7A		*Streptomyces sp*	Melibiase 2 (galactosidase)	PF16499
PropLec7A		*Frigoribacterium sp*	Arabinosidase, galactosidase	PF04616
PropLec7B		*Streptomyces davaonensi*	Peptidase S8	PF00082
PropLec7B		*Scytonema hofmannii*	Chitinase	PF00704

Occurrence of novel assembly fold for β-propeller

As described above, β-propellers are generally consisting of one peptide presenting a tandem-repeat. The only exception occurred in the PropLec6A family: these lectins have been characterized in three fungi (see Table S1) with six blade repeats for a domain approximately 300 amino acid-long, but also in bacteria with two

blade repeats in a 90 amino acid domain, that trimerizes to form the same 6-blade propeller [15, 44]. This is the only case of natural β-propeller assembled by oligomerization. The bimodal distribution of blade numbers in PropLec6A family, with maxima at 6-blade and 2-blade is shown in Figure 5 and in supplemental information (Figure S8). However, from the graph distribution, we predicted that 3-blade domains could also exist, which would correspond to a β-propeller formation by dimerization that was never observed before.

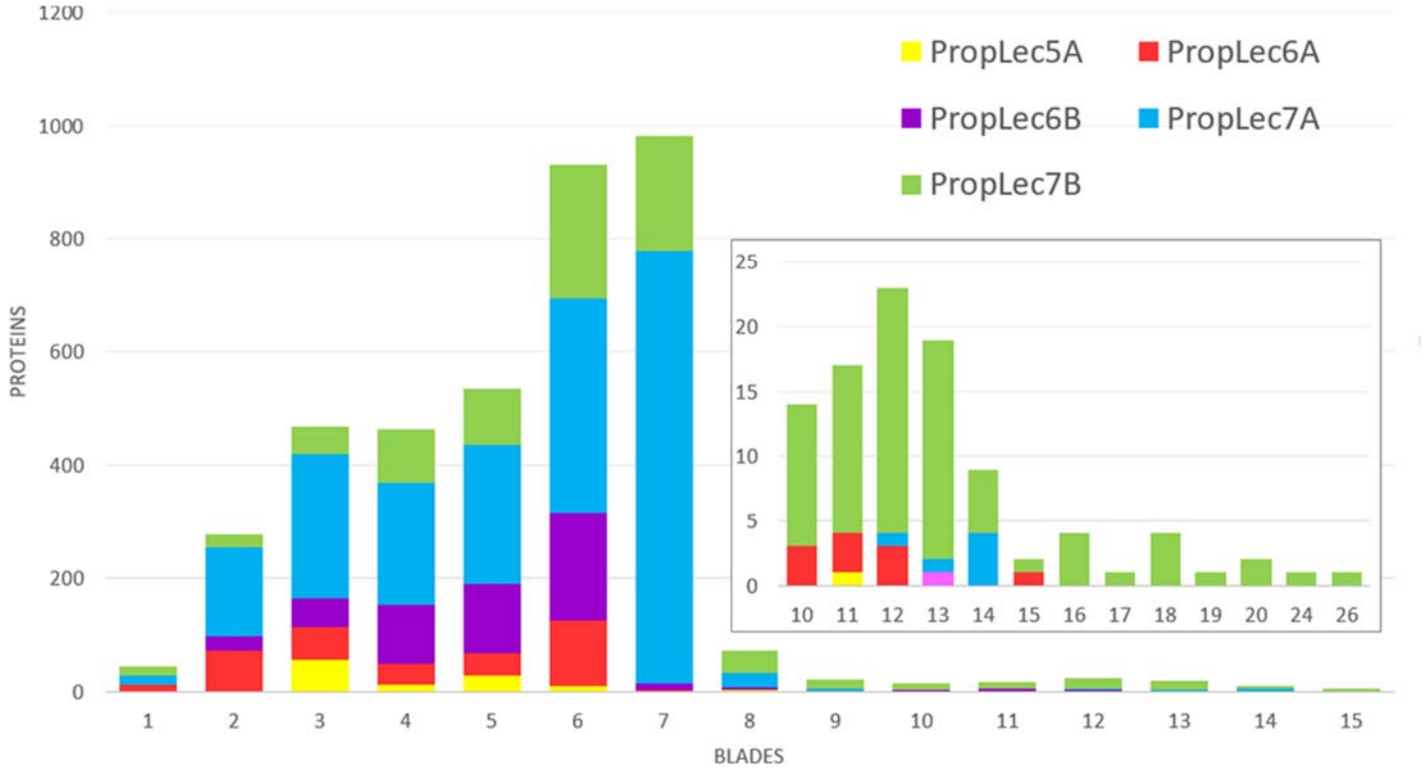

Figure 5. *Analysis of number of adjacent blades in predicted PropLecs*

The predicted 3-blade sequences of PropLec6A were therefore analysed to select those with a high similarity score, an approximate size of 150 amino acids (three repeats) and correct gene start and ending. Four sequences were selected, and annotated as 3-blades lectins : UPI0009E3DCE8 in *Kordia zhangzhouensis*[46] and A0A2T6C3M6 in *K. periserrulae*[47], bacteria from freshwater and marine environment, respectively, as well as A0A1V6N7V4 in *Penicillium polonicum* and A0A124GTL0 in *P. freii*, two filamentous fungi responsible for the production of mycotoxins[48]. The alignment of blade sequences of the *K. zhangzhouensis* lectin (KozL) and *P. polonicum* one (PepL) are displayed in Figure 6. Both proteins present conservation of all the amino acids involved in fucose binding and can therefore be annotated as putative lectins. Analysis of the identity matrix at the blade level (Figure S9) demonstrates a strong conservation of blades within the KozL sequence (55 to 62% identity), higher than in the other sequences of PropLec6A group. Internal conservation is low within PepL blade sequences (9 to 25%). Blade sequences of KozL present stronger similarity to bacterial lectin (BambL) than to fungal ones, as expected.

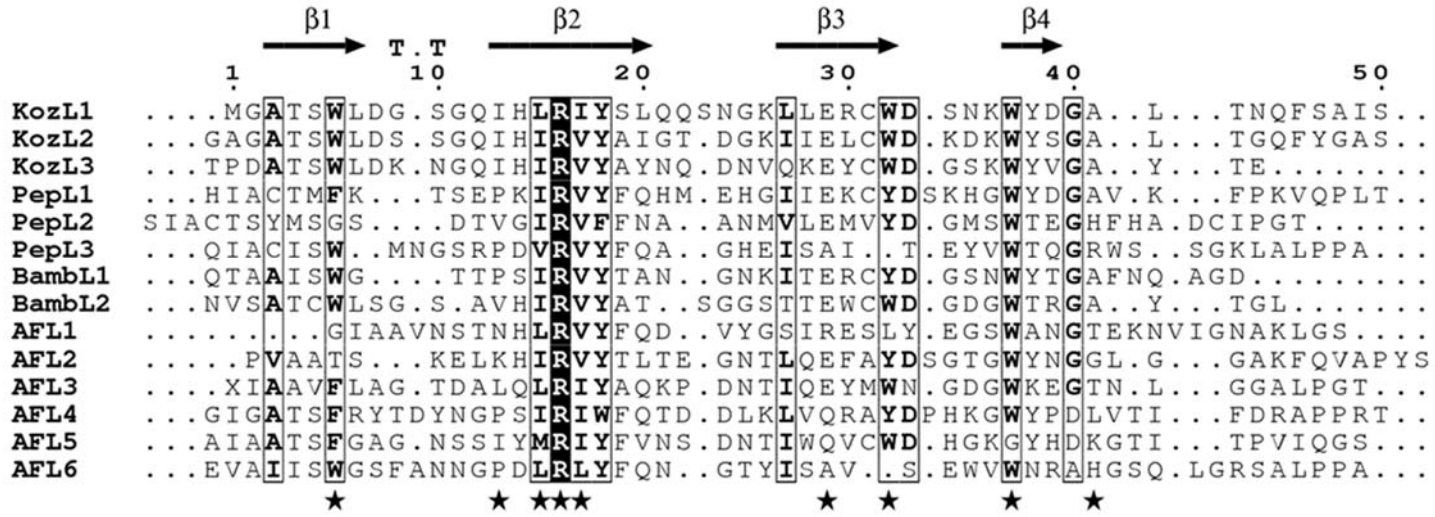

Figure 6. *Alignment of sequences for the 3-blade proteins KoZL and PepL with selected 2-blade and 6-blade members of PropLec6A family*

The genes coding for KozL and PepL were synthetized after appropriate codon optimization and expressed in *Escherichia coli.* Although PepL formed inclusion bodies, Kozl was obtained in a soluble form with expected size of 16 kDa. It was produced and purified on a carbohydrate-affinity column as previously described for RSL[15]. The protein is fully functional with very strong affinity for fucose as determined by titration microcalorimetry (figure 7A). A dissociation constant (Kd) of 0.86 µM is obtained for methyl-α-L-fucoside (MeFuc), in agreement with affinity previously measured with RSL and BambL. Titration microcalorimetry is also suited for measuring the molar ration of ligand to protein, and a value of 3.2 was obtained, confirming the presence of three active binding sites on each KozL protomer.

Crystals of KozL complexed with MeFuc were obtained by co-crystallisation. Diffraction data were collected on beam line PX1 at Soleil synchrotron to 1.55 Å resolution in $P22_12_1$ space group. Attempts to solve the structure by molecular replacement method were not successful. A methyl-α-L-selenofucoside derivative (SeFuc), synthetized as previously described[15], was cocrystallised for SAD phasing and data were collected at 2.65 Å resolution. Statistics for both complexes are described in Table S2 (sup. Info), and only the structure of KozL with αMeFuc was fully refined and described herein.

The asymmetric unit contains four monomers of KozL assembled in two β-propellers, and two additional monomers that form another dimer of β-propeller when applying the 2-fold symmetry of the space group. The tetramer formed by chains ABCD (Figure 7C) presents an interface of 13 000 $Å^2$ as calculated by PISA (PDBe.org). The oligomeric state in solution was confirmed by analytical ultracentrifugation (Figure S10). The dominant peak of KozL (90% of the total signal) has a sedimentation coefficient of 4.4 S (4.6 S at standard conditions) and corresponds to the tetrameric species with a moderately elongated shape (f/f0 = 1.3).

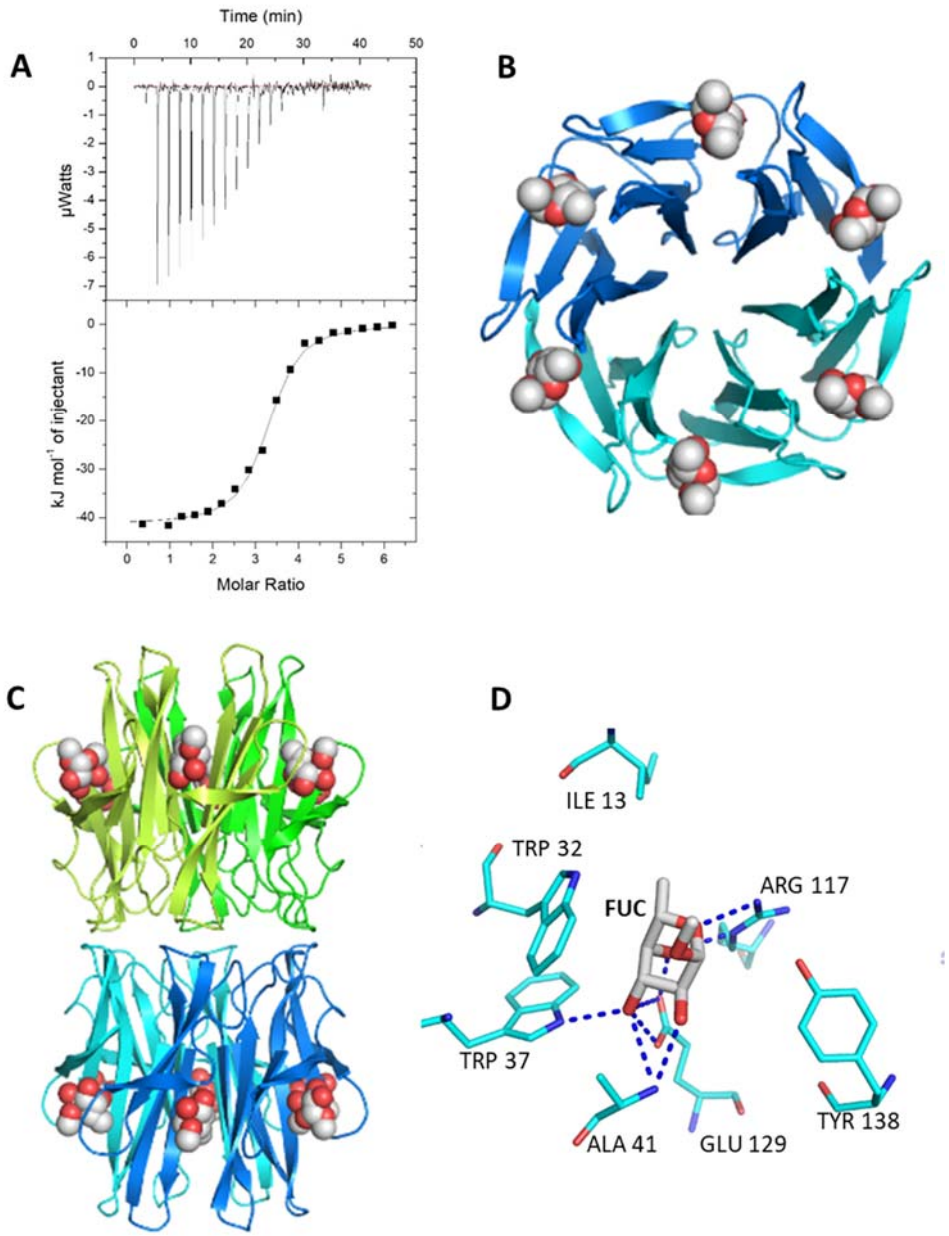

Figure 7. *Function and structure of KozL. A. ITC data with thermogram (top) and integrated peaks (bottom), B. Dimer of KozL assembles in a β-propeller structure with each chain represented by a shade of blue, and MeFuc ligand represented by spheres. C. Tetramer of KozL assembled in dimeric association of β-propellers. D. The fucose binding site in one of the three binding sites with hydrogen bonds is represented by blued dashed lines.*

Electron density clearly indicates the presence of 18 MeFuc residues (three per monomers), with few additional molecules of crystallizing agents (2-methyl-2,4-pentanediol, ethanediol and nonaethylene glycol) and water molecules. The binding sites are located between the blades, with two intramolecular and one intermolecular sites. The amino acids involved in fucose binding are fully conserved in the three blades and are very similar to what has been observed in BambL and RSL. Fucose is stabilized by hydrogen bonds to side chains of Arg (16/67/117 for the three sites), Glu (29/79/129) and Trp (87/137/37*) and to main chain of Ala (41/91/141) and by hydrophobic interactions with another Trp indol ring (82/132/32*) and Ile (64/114/13*) (Figure 7D).

Discussion and Conclusion

The blade signatures that have been designed in the present study allowed for the identification of new PropLec sequences in a wide collection of genomes. The KozL protein, that was not annotated previously as a lectin, provided an experimental validation of our approach. The protein function was confirmed by measuring its strong affinity for fucose. Apart from the conservation of binding sites and the 4-strand β-sheet repeats, KozL is rather different from other PropLec6A structures, especially the loops on both side of the donuts (Figure 8A). The β-propeller of KozL is formed by dimerization of two 3-blade domains, which has never been reported before and is of high interest in the evolution of proteins. As illustrated in Figure 8B, in the PropLec6A family, the donut shape of the β-propeller can be formed by dimerization (KoZL), trimerization (for bacterial lectins BambL/RSL) or can be monomeric (for fungal lectins AFL/AAL/AOL). Evolution used symmetry in a very efficient way to build the same objects from different numbers of domain repeats.

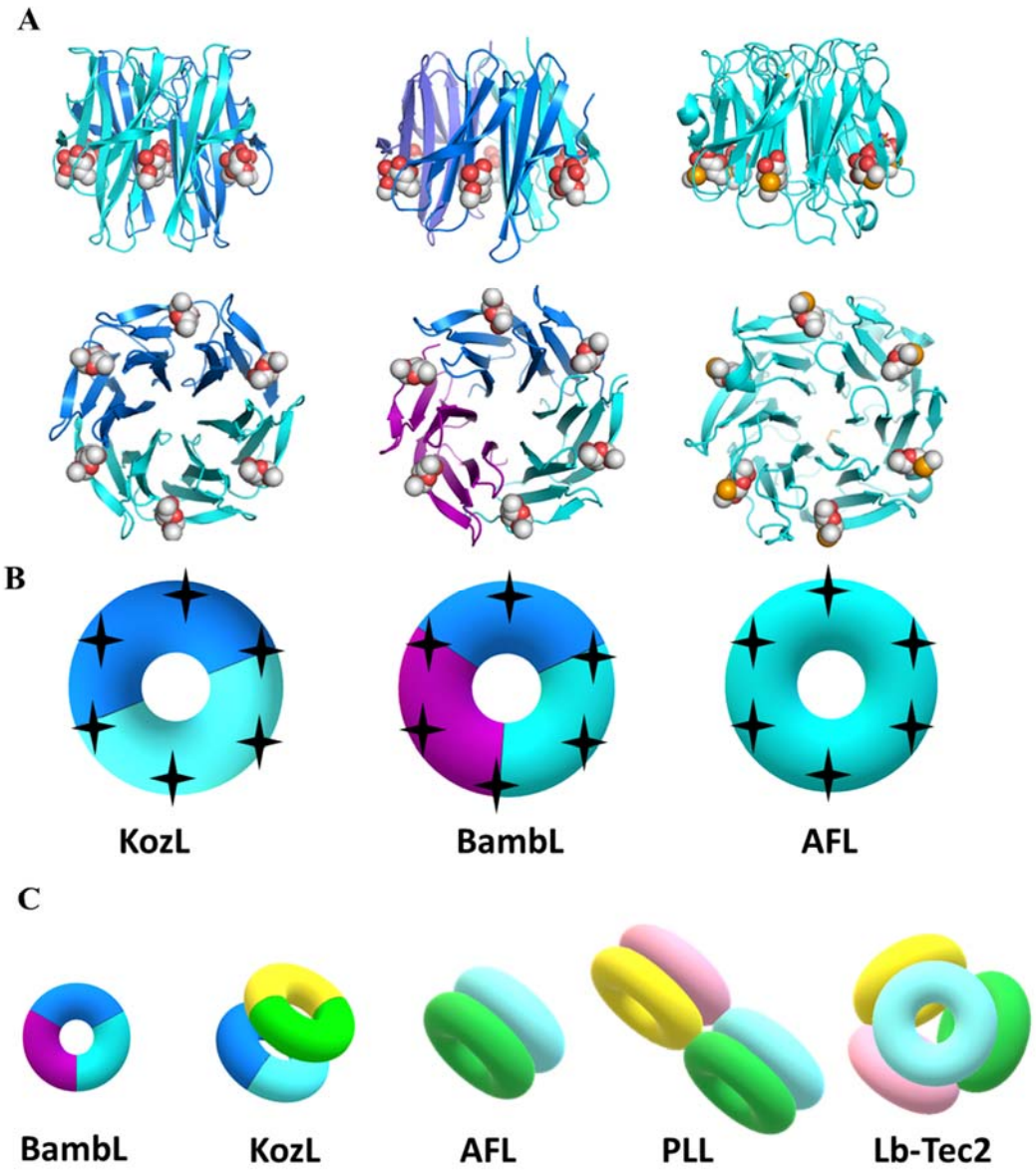

Figure 8. *A. Different oligomerisation modes for the creation of6-propeller structure in PropLec6A family. In the donut schematic representation, the stars denote glycan binding sites. B. Different assemblies of β-propellers observed in PropLec 3D-structures.*

Furthermore, β-propeller lectins have the ability to form supra-molecular assemblies by oligomerisation of the donut shapes, resulting in the different organisation of carbohydrate binding sites in space. Figure 8C schematizes the different oligomerization modes that have been observed so far. Some PropLecs such as BambL in the PropLec6A family, but also PVL in the PropLec7B family, occur as single β-propeller in solution, while others, such as KoZL and AFL (PropLec6A) and PHL (PropLec7A) are in the form of back-to-back propellers, that present binding sites in opposite directions. The tetrameric association of β-propellers is observed in PLL (PropLec7A), with stabilization by disulphide bridges, and in fungal tectonin Lb-Tec2 (PropLec6B) where four β-propellers form a round-shaped virus-like assembly with 24 carbohydrate binding sites evenly partitioned on the surface.

In this study, we identified almost 4000 sequences of putative PropLecs and we validated our approach with the experimental characterization of a novel structure with strong interest for evolution. Clearly, the wealth of new sequences identified opens the way to research on the evolution of β -propeller folds. Furthermore, the donut shape of PropLecs is a very robust protein structure that can be used as scaffold for building multivalent protein structures and PropLecs from pathogenic organisms are likely to be involved in host-glycan recognition and can be used as target of anti-infectious compounds.

Acknowledgements. The authors acknowledge support by the ANR PIA Glyco@Alps (ANR-15-IDEX-02) and the Alliance Campus Rhodanien Co-Funds (http://campusrhodanien.unige-cofunds.ch). A.K., M.W. and A.I. are grateful for the support of EEC Bison project (H2020-TWINN-2015-692068) for financing the stay of A.K. in Grenoble. A.K. and M.W. acknowledge the CEITEC 2020 project (LQ1601) from MEYS CR. We acknowledge the CF Biomolecular Interactions and Crystallization supported by the CIISB research infrastructure (LM2015043 funded by MEYS CR) for their support with obtaining AUC data. We are grateful to synchrotron SOLEIL (Saint Aubin, France) for access and technical support at beamline PROXIMA 1 and for the help of Pierre Legrand.

Author contributions. F.B. developed the database and build the interface under the guidance of F.L., A.I. and S.P. A. K. produced the lectin and characterized it under guidance from A.V. M.L. synthetized the selenoligand. A.V. solved the crystal structure and refined it. M.W. performed and analysed ultracentrifugation experiments. F.B., A.I and F.L. wrote the manuscript and prepared figures with the critical input of S.P., A.V. and M.W.

References

1. Lis, H. & Sharon, N. Lectins: Carbohydrate-specific proteins that mediate cellular recognition. *Chem. Rev.* **98**, 637-674 (1998).

2. Bonnardel, F., Mariethoz, J., Salentin, S., Robin, X., Schroeder, M., Perez, S., Lisacek, F. & Imberty, A. UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acids Res* (in press).

3. Chen, C.K., Chan, N.L. & Wang, A.H. The many blades of the beta-propeller proteins: conserved but versatile. *Trends Biochem Sci* **36**, 553-561 (2011).

4. Fulop, V. & Jones, D.T. Beta propellers: structural rigidity and functional diversity. *Curr Opin Struct Biol* **9**, 715-721 (1999).

5. Jawad, Z. & Paoli, M. Novel sequences propel familiar folds. *Structure* **10**, 447-454 (2002).

6. Chaudhuri, I., Soding, J. & Lupas, A.N. Evolution of the beta-propeller fold. *Proteins* **71**, 795-803 (2008).

7. Kopec, K.O. & Lupas, A.N. beta-Propeller blades as ancestral peptides in protein evolution. *PLoS One* **8**, e77074 (2013).

8. Beisel, H.G., Kawabata, S., Iwanaga, S., Huber, R. & Bode, W. Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus. *EMBO J* **18**, 2313-2322 (1999).

9. Kawabata, S. & Iwanaga, S. Role of lectins in the innate immunity of horseshoe crab. *Dev Comp Immunol* **23**, 391-400 (1999).

10. Wimmerova, M., Mitchell, E., Sanchez, J.F., Gautier, C. & Imberty, A. Crystal structure of fungal lectin: Six-bladed b-propeller fold and novel recognition mode for *Aleuria aurantia* lectin. *J. Biol. Chem.* **278**, 27059-27067 (2003).

11. Houser, J. et al. A soluble fucose-specific lectin from *Aspergillus fumigatus* conidia - Structure, specificity and possible role in fungal pathogenicity. *PloS ONE* **8**, e83077 (2013).

12. Kerr, S.C. et al. FleA Expression in Aspergillus fumigatus Is Recognized by Fucosylated Structures on Mucins and Macrophages to Prevent Lung Infection. *PLoS Pathog* **12**, e1005555 (2016).

13. Richard, N. et al. Human bronchial epithelial cells inhibit *Aspergillus fumigatus* germination of extracellular conidia via FleA recognition. *Scientific Reports* **8**, 15699 (2018).

14. Audfray, A. et al. A recombinant fungal lectin for labeling truncated glycans on human cancer cells. *PLoS One* **10**, e0128190 (2015).

15. Kostlanová, N., Mitchell, E.P., Lortat-Jacob, H., Oscarson, S., Lahmann, M., Gilboa-Garber, N., Chambat, G., Wimmerová, M. & Imberty, A. The fucose-binding lectin from *Ralstonia solanacearum*: a new type of -propeller architecture formed by oligomerisation and interacting with fucoside, fucosyllactose and plant xyloglucan. *J. Biol. Chem.* **280**, 27839-27849 (2005).

16. Capaldi, S., Faggion, B., Carrizo, M.E., Destefanis, L., Gonzalez, M.C., Perduca, M., Bovi, M., Galliano, M. & Monaco, H.L. Three-dimensional structure and ligand-binding site of carp fishelectin (FEL). *Acta Crystallogr D Biol Crystallogr* **71**, 1123-1135 (2015).

17. Sommer, R., Makshakova, O.N., Wohlschlager, T., Hutin, S., Marsh, M., Titz, A., Kunzler, M. & Varrot, A. Crystal structures of fungal tectonin in complex with O-methylated glycans suggest key role in innate immune defense. *Structure* **26**, 391-402 (2018).

18. Wohlschlager, T. et al. Methylated glycans as conserved targets of animal and fungal innate defense. *Proc Natl Acad Sci U S A* **111**, E2787-2796 (2014).

19. Cioci, G. et al. b-Propeller crystal structure of *Psathyrella velutina* lectin: An integrin-like fungal protein interacting with monosaccharides and calcium. *J. Mol. Biol.* **357**, 1575-1591 (2006).

20. Ren, X.M., Li, D.F., Jiang, S., Lan, X.Q., Hu, Y., Sun, H. & Wang, D.C. Structural basis of specific recognition of non-reducing terminal N-acetylglucosamine by an *Agrocybe aegerita* Lectin. *PLoS One* **10**, e0129608 (2015).

21. Jancarikova, G., Houser, J., Dobes, P., Demo, G., Hyrsl, P. & Wimmerova, M. Characterization of novel bangle lectin from *Photorhabdus asymbiotica* with dual sugar-binding specificity and its effect on host immunity. *PLoS Pathog* **13**, e1006564 (2017).

22. Kumar, A., Sykorova, P., Demo, G., Dobes, P., Hyrsl, P. & Wimmerova, M. A novel fucose-binding lectin from P*hotorhabdus luminescens* (PLL) with an unusual heptabladed beta-propeller tetrameric qtructure. *J Biol Chem* **291**, 25032-25049 (2016).

23. Goyard, D., Baldoneschi, V., Varrot, A., Fiore, M., Imberty, A., Richichi, B., Renaudet, O. & Nativi, C. Multivalent glycomimetics with affinity and selecivity towards fucose-binding receptors from emerging pathogens. *Bioconjugate Chemistry* **29**, 83−88 (2018).

24. Jancarikova, G., Herczeg, M., Fujdiarova, E., Houser, J., Kover, K.E., Borbas, A., Wimmerova, M. & Csavas, M. Synthesis of alpha-l-Fucopyranoside-Presenting Glycoclusters and Investigation of Their Interaction with Photorhabdus asymbiotica Lectin (PHL). *Chemistry* (2018).

25. Machida, T., Novoa, A., Gillon, É., Zheng, S., Claudinon, J., Eierhoff, T., Imberty, A., Römer, W. & Winssinger, N. Dynamic cooperative glycan assembly blocks binding of bacterial lectins to epithelial cells *Angewandte Chemie International Edition* **56**, 6762-6766 (2017).

26. Liu, W. et al. AANL (Agrocybe aegerita lectin 2) is a new facile tool to probe for O-GlcNAcylation. *Glycobiology* **28**, 363-373 (2018).

27. Machon, O., Baldini, S.F., Ribeiro, J.P., Steenackers, A., Varrot, A., Lefebvre, T. & Imberty, A. Recombinant fungal lectin as a new tool to investigate O-GlcNAcylation processes. *Glycobiology* **27**, 123-128 (2017).

28. Arnaud, J. et al. Reduction of lectin valency drastically changes glycolipid dynamics in membranes, but not surface avidity. *ACS Chemical Biology* **8**, 1918-1924 (2013).

29. Arnaud, J., Tröndle, K., Claudinon, J., Audfray, A., Varrot, A., Römer, W. & Imberty, A. Membrane deformation by neolectins with engineered glycolipid binding sites. *Angewandte Chemie International Edition* **53**, 9267–9270 (2014).

30. Yadid, I. & Tawfik, D.S. Reconstruction of functional beta-propeller lectins via homo-oligomeric assembly of shorter fragments. *J Mol Biol* **365**, 10-17 (2007).

31. Yadid, I. & Tawfik, D.S. Functional beta-propeller lectins by tandem duplications of repetitive units. *Protein Eng Des Sel* **24**, 185-195 (2011).

32. Dawson, N.L., Sillitoe, I., Lees, J.G., Lam, S.D. & Orengo, C.A. CATH-Gene3D: Generation of the resource and its use in obtaining structural and functional annotations for protein sequences. *Methods Mol Biol* **1558**, 79-110 (2017).

33. Finn, R.D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**, D279-285 (2016).

34. Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A. & Eddy, S.R. HMMER web server: 2015 update. *Nucleic Acids Res* **43**, W30-38 (2015).

35. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H. & UniProt, C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).

36. Heger, A. & Holm, L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins* **41**, 224-237 (2000).

37. Maguire, E., Rocca-Serra, P., Sansone, S.-A. & Chen, M. in Eurographics Conference on Visualization (EuroVis). (eds. N. Elmqvist, M. Hlawitschka & J. Kennedy) (2014).

38. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

39. Brown, G.R. et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res* **43**, D36-42 (2015).

40. O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).

41. Low, D.H., Ang, Z., Yuan, Q., Frecer, V., Ho, B., Chen, J. & Ding, J.L. A novel human tectonin protein with multivalent beta-propeller folds interacts with ficolin and binds bacterial LPS. *PLoS One* **4**, e6260 (2009).

42. Imberty, A. in E-book : Synthesis and Biological Applications of Glycoconjugates. (eds. O. Renaudet & N. Spinelli) 3-11 (Bentham Science Publishers Ltd, 2011).

43. Sharon, N. Carbohydrates as future anti-adhesion drugs for infectious diseases. *Biochim Biophys Acta* **1760**, 527-537 (2006).

44. Audfray, A. et al. The fucose-binding lectin from opportunistic pathogen *Burkholderia ambifaria* binds to both plant and human oligosaccharidic epitopes. *Journal of Biological Chemistry* **287**, 4335-4347 (2012).

45. Boraston, A.B., Bolam, D.N., Gilbert, H.J. & Davies, G.J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem J* **382**, 769-781 (2004).

46. Du, J., Liu, Y., Lai, Q., Dong, C., Xie, Y. & Shao, Z. Kordia zhangzhouensis sp. nov., isolated from surface freshwater. *Int J Syst Evol Microbiol* **65**, 3379-3383 (2015).

47. Choi, A., Oh, H.M., Yang, S.J. & Cho, J.C. Kordia periserrulae sp. nov., isolated from a marine polychaete Periserrula leucophryna, and emended description of the genus Kordia. *Int J Syst Evol Microbiol* **61**, 864-869 (2011).

48. Mills, J.T., Seifert, K.A., Frisvad, J.C. & Abramson, D. Nephrotoxigenic Penicillium species occurring on farm-stored cereal grains in western Canada. *Mycopathologia* **130**, 23-28 (1995).

Material and Methods (for on line access)

Database construction

The features, taxonomy and identified domains of the predicted lectins (from UniRef100 database release of 12/09/18 with HMMSEARCH version 3.2) are stored in distinct tables to preserve the reactivity of the web platform and avoid computing information on the run. Predicted protein information was collected from UniProt and corresponding RefSeq entry including data on the related 1889 species. 18545 Pfam domains from 194 Pfam families have also been identified on the predicted proteins (from PFAM-A release of 15/10/18). The information is fetched using the predicted protein UniProt AC and PYTHON 3 scripts and the information files are loaded in the database with PHP scripts to facilitate maintenance and update.

Web module construction

The UniLectin web platform (https://www.unilectin.eu) is dedicated to the classification and curation of lectin structures (UniLectin3D module) and prediction of lectin sequences in genomes. The module dedicated to β-popeller lectins (PropLec) is available on the UniLectin platform. The interface has been developed with PHP version 7, Bootstrap version 3, and MySQL database version 5.6. Interactive graphics are developed in JavaScript based on D3JS libraries version 3 and dynamically generated to match the research criteria selected by the user.

Defining a similarity score

HMMER default scores is not comparable between predicted proteins with a different number of blades. To avoid the bias we defined a new similarity score. We use the alignment of the reference seed and the predicted blades performed with MUSCLE to define a quality score for each predicted protein. The similarity score is shown below. To control the bias due to variable numbers of predicted blades and different lengths of conserved blade domain in distinct families, the calculations are centered on those two criteria.

$$score = \sum_{i}^{1:len(MSA)} REF\_FREQ[MAX(AAi)] * PRED\_FREQ[MAX(AAi)]$$

$$Sim_{score} = \frac{score - MEAN(scores\_family\_nbblades)}{SD(scores\_family\_nbblades)}$$

Where REF_FREQ is the frequency of the most frequent amino acid (MAX (AAi)) at position i in the reference/seed domain and PRED_FREQ is the frequency of the most frequent amino acid at position i in the predicted protein. Score distributions by family are shown in Figure S7

Cloning of Kum and Kordia genes

The peptide sequences of the putative lectins KozL from *Kordia zhangzhouensis* (UPI0009E3DCE8) and PepL from *Penicillium polonicum* (A0A1V6N7V4) were translated into nucleotide sequence and were synthesized after codon optimization for expression in Escherichia coli (Eurofins Genomics,Germany). The genes were introduced in the pET-TEV expression vector using NcoI and XhoI restriction sites and enzymes (New England Biolabs)[1]. The pET-TEV-KozL and pET-TEV-PepL vectors were transformed into *E. coli* BL21(DE3)

Production and purification of KoZL

E. coli BL21 (DE3) cells harboring the plasmid pET-TEV-KozL were cultured in LB Broth medium with 30 $\mu g.mL^{-1}$ kanamycin at 37°C. When the culture reached an A600nm of 0.6-0.8, protein expression was induced with 0.1 mM isopropyl β-D-thiogalactoside. After 3 hours at 37 °C, cells were harvested by centrifugation at 6000 x g for 10 min and frozen at -20 °C. The pellet from 1 liter culture was resuspended in 30 mL of buffer A composed of 30 mM Tris-HCl pH 8.5 and 500mM NaCl prior addition of 1µl of Benzonase endonucluease (Sigma-Aldrich). After 15 min incubation at room temperature, cells were disrupted at a pressure of 1.9 kBar (Constant Cell Disruption System). The cell lysate was centrifuged at 24000 x g for 30 min at 4°C and the supernatant was filtered on 0.45 µm prior loading on a 10 ml mannose agarose column (Sigma-Aldrich) pre-equilibrated with buffer A. The column was washed with buffer A to remove unbound proteins and elution was performed with buffer A supplemented with 20 mM mannose. Purity was checked on 15% SDS-PAGE gel (15 %) before pooling of the appropriate fractions for dialysis with buffer B composed of 20 mM Tris pH 8.5, 250 mM NaCl. KozL was concentrated by centrifugation using a Vivaspin (3KDa, Sartorius) and stored in fridge for further use. For the long-term storage at -20°C, KozL was dialysed against ultrapure water, and lyophilised and kept in deep fridge. The total yield of purified KozL was 35 mg from 4.5 g of cells. All expression conditions tested for pET-TEV-KozL led to the formation of inclusion bodies to date.

ITC experiments

ITC experiments were performed with isothermal titration calorimeters (MicroCaliTC200; Malvern). Experiments were carried out at 25 ºC ± 0.1 ºC. Methyl-α-fucoside (TCI) solution was prepared in same buffer as KozL. The ITC cell contained 0.02 mM mM of KozL and the syringue 0.6 mM of MeFuc. The ligand was added by infection of 2 µL at intervals of 2 min while stirring at 1000 rpm. Prior to sample analysis, a control experiment, where the protein sample in the calorimeter cell was substituted by buffer, was performed, resulting in insignificant heat of dilution. Integrated heat effects were analysed by nonlinear regression using a one site binding model (Microcal Origin 7). The experimental data fitted to a theoretical titration curve gave the association constant Ka and the enthalpy of binding (ΔH). The experiments were performed in duplicates

Analytical ultracentrifugation experiments

Analytical ultracentrifugation experiments were performed using ProteomeLab XL-I analytical ultracentrifuge (Beckman Coulter) equipped with An-60 Ti rotor. Before analysis, lyophilized KoZL was

dissolved in the experimental buffer (20 mM Tris, 150 mM NaCl, pH 7.4) and the buffer was used as an optical reference.

Sedimentation velocity experiments were conducted in titanium double-sector centerpiece cells (Nanolytics Instruments, Germany) loaded with 380 µL of both protein sample (0.02-0.17 $mg.mL^{-1}$) and reference solution. Data were collected using absorbance optics at 20 °C at a rotor speed of 50,000 rpm. Scans were performed at 280 nm at 4 min intervals and 0.003 cm spatial resolution in continuous scan mode. The partial specific volume of protein and the solvent density and viscosity were calculated from the amino acid sequence and buffer composition, respectively, using the software Sednterp (http://bitcwiki.sr.unh.edu). The sedimentation profiles were analyzed with the program Sedfit 15.01 [2]. Continuous c(s) distribution model was used for the analysis.

Crystallization and structure determination of KozL

Crystallization experiments were performed using the hanging-drop vapor-diffusion method with drops made of 1 µL of protein at 10 $mg.mL^{-1}$ in buffer B and 1 µL of reservoir solution at 19 °C. Commercial screens (Morpheus I and II, Clear Strategy Screen I and II and BCS; Molecular Dimensions Ltd) led to several crystallization hits. Cocrystallisation with 20 mM MeFuc using the solution 1-40 from Morpheus 1[3] (120 mM alcohols, 100mM buffer pH 6.5, 37.5% MPD/PEG1000/PEG3350) results in parallelepiped crystals in 3-5 days. Thick rods were obtained after cocrystallisation of KozL incubated with 1 mM methyl-α-selenofucoside (SeFuc)[4] in 35% PEG smear medium, 10% isopropanol optimized from hit from solution 2-38 of the BCS screen[5]. Crystals were directly mounted in a Litholoop (Molecular Dimensions Ltd) and flashed freezed in liquid nitrogen. Data were collected using a Pilatus 6M detector (Dectris Ltd) on the Proxima-1 beamline at SOLEIL, Saint Aubin, France. The data were processed using XDS[6]. All further computing was performed using the CCP4 suite and interfaces[7] (Table S2).

Structural determination and refinement. The structure of KozL was solved by SAD method at the selenium peak (λ = 0.97914 Å) using the signal of the selenated ligand. ShelXC/D[8] found 21 selenium sites with CC-All 34.09 and CFOM of 52.35. Since phases obtained by those sites were not of good enough quality to allow hand determination and initial model building using with ShelxE, 10 of the selenium sites related by non-crystallographic operator as determined using Profess were used for SAD phasing using heavy metal site in PHASER[9]. Density modification was then performed using Parrot[10] and initial model building with Buccaneer[11]. Only protein chains with assigned sequence were then used for molecular replacement of the complex data of KozL in complex with Mefuc at 1.55 Å using Phaser. Initial autobuilding and refinement of the 6 protein chains was performed with Buccaneer followed by iterative structure refinement with Refmac5.8[12] and manual model corrections in COOT[13]. 5% of the observations were set aside for cross-validation analysis and riding hydrogen atoms were added and used for geometry and structure-factor calculations. The stereochemical quality of the refined models was validated on the wwPDB Validation server:

http://wwpdb-validation.wwpdb.org and carbohydrates were checked in Privateer[14]. All figures were drawn with PyMOL Molecular Graphic System program (Version 2.0.4, Schrodinger, LLC).

Accession codes. Coordinates of the structure of KozL in complex with MeFuc and structure factors for both MeFuc and MeSeFuc complex data have been deposited in the Protein Data Bank (https://www.rcsb.org/)[15] under accession codes 6HTN.

References

1. Houben, K., Marion, D., Tarbouriech, N., Ruigrok, R.W. & Blanchard, L. Interaction of the C-terminal domains of sendai virus N and P proteins: comparison of polymerase-nucleocapsid interactions within the paramyxovirus family. *J Virol* **81**, 6807-6816 (2007).
2. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys J* **78**, 1606-1619 (2000).
3. Gorrec, F. The MORPHEUS protein crystallization screen. *J Appl Crystallogr* **42**, 1035-1042 (2009).
4. Kostlanová, N., Mitchell, E.P., Lortat-Jacob, H., Oscarson, S., Lahmann, M., Gilboa-Garber, N., Chambat, G., Wimmerová, M. & Imberty, A. The fucose-binding lectin from *Ralstonia solanacearum*: a new type of -propeller architecture formed by oligomerisation and interacting with fucoside, fucosyllactose and plant xyloglucan. *J. Biol. Chem.* **280**, 27839-27849 (2005).
5. Chaikuad, A., Knapp, S. & von Delft, F. Defined PEG smears as an alternative approach to enhance the search for crystallization conditions and crystal-quality improvement in reduced screens. *Acta Crystallogr D Biol Crystallogr* **71**, 1627-1639 (2015).
6. Kabsch, W. Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132 (2010).
7. Winn, M.D. et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235-242 (2011).
8. Schneider, T.R. & Sheldrick, G.M. Substructure solution with SHELXD. *Acta Crystallogr D Biol Crystallogr* **58**, 1772-1779 (2002).
9. McCoy, A.J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr D Biol Crystallogr* **63**, 32-41 (2007).
10. Cowtan, K. Recent developments in classical density modification. *Acta Crystallogr D Biol Crystallogr* **66**, 470-478 (2010).
11. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr D Biol Crystallogr* **62**, 1002-1011 (2006).
12. Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F. & Vagin, A.A. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355-367 (2011).
13. Emsley, P., Lohkamp, B., Scott, W.G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).
14. Agirre, J., Iglesias-Fernandez, J., Rovira, C., Davies, G.J., Wilson, K.S. & Cowtan, K.D. Privateer: software for the conformational validation of carbohydrate structures. *Nat Struct Mol Biol* **22**, 833-834 (2015).
15. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).